原著論文

形態素解析を用いた事故・リスクに関する記述の分析

湊 淳, 鈴木 慎吾, 村上 雄太郎, 伊多波 正徳, 小澤 哲

茨城大学

Research on Texts of Accidents and Risks using Morphological Analysis

Atsushi MINATO, Shingo SUZUKI, Yutaro MURAKAMI, Masanori ITABA and Satoru OZAWA

Ibaraki University, 4-12-1 Nakanarusawa, Hitachi-shi, Ibaraki 316-8511, Japan

Abstract : In this research, we applied morphological analysis to abstract the psychological information of writer. Chasen is a Japanese software for morphological analysis. Two types of texts are evaluated. (1) Explanation of the outline of well-known accident or trouble. (2) Author's impression about the accident or trouble. Frequencies of verb, adjective, adverb, Rentaishi, conjunction, particle, auxiliary verb and exclamation are compared in both texts. It was found that the frequency of adverb is higher in the author's impressions. The tf-idf of adverb was also examined. it was found that the tf-idfs of adverbs belong to two groups are significantly different.

Keywords : *Morphological analysis, Chasen, psychological condition, text analysis, Japanese language, risk, trouble, tf-idf*

1. はじめに

インターネット上には、個人が書き込んだテキストデータが膨大な数存在する。電子掲示板サイト、ソーシャル・ネットワークワーキング・サービス、商品の比較サイトなどに、原子力への不安、購買した製品への感想など、さまざまな個人の意見が含まれている。膨大な文献の中から、特定の感情が含まれる文章群を抜き出すことができれば、政策決定、マーケティングなどに活用できる。

文献データから著者に関する情報を抽出し、真贋分析などを行う方法に関しては、数多くの研究が行われている [1, 2]。またアンケート、コールセンターに寄せられた意見などのテキストに対して、統計処理などを行い、意味のある情報を取り出す方法は、テキストマイニングとして知られている [3-7]。

西洋の文章の場合、単語がスペースで区切られているので、活用変化を含む辞書データがあれば、自動解析が可能である。コンピュータで日本語の文章を自動的に処理する場合の問題点は、最初に文章を品詞ごとに分割する作業が必要な点にある。近年になり、実用可能な精度で形態素解析を行うソフトウェアが公開されるようになった。形態素解析のソフトウェアを使うと日本語の文章を品詞ごとに分類し、品詞の分類まで自動で行うことができる。

本研究では、「怒り」などの感情を伴った文章のふり分けを目的として、品詞の出現頻度に関する基礎的評価を行った。名詞や動詞はテキストマイニングでキーワードとして用いられることが多い。本研究では、動詞などを修飾し、話し手の気持ちを表すと考えられる副詞に着目して分析を行った。

今回は、同じ著者がリスクやトラブルに関して書いた文章において、事実の概要に関する記述と、感想文を対象として品詞の出現頻度、情報検索でよく利用される tf-idf の値がどのように変化するか調べた。副詞に関しては、概要と感想文の tf-idf の値に有意な差があることが分かった。

2. 事故・トラブルに対する記述の解析

本研究では茶筌を用いた形態素解析により、各品詞の出現頻度と tf-idf の値が著者の感情にどのように依存するかを調べた。

形態素解析には、奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座で開発された「茶筌 (Chasen) Version2.3.3」を用いた [8, 9]。茶筌システムの原形は、京都大学長尾研究室および奈良先端科学技術大学院大学情報科学研究科において開発された日本語形態素解析システム JUMAN (version2.0) である。テキスト形式の入力ファイルを用意して、出力オプションを指定して、茶筌を実行することにより、品詞によって分けられたデータが得られる。

70名の理工系修士課程の1年生(全員男性)に、表1に示すような事故、企業トラブルなどに関するレポートを書いてもらった。課題に関しては、良く知られていて、参考文献やインターネット上の資料も数多くあり、組織などの不安全行動、安全管理の不備などが原因となるものを中心に選んだ。同じ課題に学生が集中しないように、くじ引きの形式で学生に一人1課題を割り当てた。複数の課題を用意した理由は、学生が他人の意見を参考にしないようにするため、また1つの課題に絞ると、その課題の固有の問題に記述が依存してしまう可能性があるためである。

表1 国内外の事故・トラブル

米スペースシャトル「コロンビア」の空中分解
もんじゅ事故
チェルノブイリ事故
スリーマイル島原子力発電所の事故
森永砒素ミルク事件
雪印乳業食中毒事件
日本航空・東京-大阪123便墜落事故
「えひめ丸」衝突事故の概要
三菱自動車工業のリコール隠し問題
1999年H2ロケット事故
1998H-IIロケット5号機の事故
日本ハムグループによる牛肉偽装問題
ナホトカ号重油流出事故
薬害エイズ事件（ミドリ十字）
カイワレO157訴訟
みずほ銀行システム障害
京福電気鉄道越前本線列車衝突事故
豊浜トンネル崩落事故
新宿・歌舞伎町ビル火災
静岡駅前地下街ガス爆発火災
東名高速道路日本坂トンネル車両火災
エンロン不正会計問題
東芝ビデオ修理クレーム事件
三菱重工発電用タービン・ローター爆発事故
サリドマイド事件
スモン事件
出光興産北海道製油所火災
2003米とカナダの大規模停電

1ヶ月ほどの時間を与え、インターネットや文献を調べてまとめるように指示した。レポートをまとめる際に、概要と感想文を分けて書いてもらった。概要は、原因、経過、社会的影響などについて最低800字で書いてもらった。感想文の方は、事故・トラブルに対する学生本人の感想とし、最低500字で書いてもらった。文字数の指定に関しては、概要に関しては資料などを参考にしている程度の量が書けるが、意見、感想を書くのは苦手な学生が多いため少ない文字数を指定した。

以下にレポートの例を示す。

・雪印乳業食中毒事件の概要

「平成12年6月末に発生した雪印乳業大阪工場（以下「大阪工場」）製造の低脂肪乳等を原因とする食中毒事件の有病者数は、14849名（平成12年9月8日現在）に達し、近年例をみない大規模食中毒事件となった。平成12年7月2日、低脂肪乳から黄色ブドウ球菌のエンテロトキシンA型が検出されたことから、大阪市は、これを病因物質とする食中毒と断定して、大阪工場を営業禁止とした。（以下略）」

・雪印乳業食中毒事件の感想

「この雪印乳業食中毒事件で私が恐ろしく思うことは、老若男女誰もが普段口にする乳製品によって食中毒が起きたということである。「イライラしているからカルシウムを取ろう」「子供の成長にはカルシウムが必要不可欠」といったことは誰もが一度は思ったことのあることで、それによってカルシウム摂取の一番効率のいい食べ物として誰もが思い浮かべるものが、『牛乳』や『ヨーグルト』などの乳製品であろう。（以下略）」

・東名高速道路日本坂トンネル車両火災の概要

「1979年（昭和55年）7月11日午後6時38分頃、静岡県焼津市と静岡市の境の東名高速道路日本坂トンネル内の下り線で、大型トラックと乗用車系5台が玉突き衝突を起こした。燃料のガソリンが火花により引火爆発、さらにトラック積荷の合成樹脂、揮発性油等が炎上し、トンネル内で数珠つなぎになった車両が次々と炎上し、死者7名、負傷者3人、車両173台を焼失する、道路のトンネル内の事故としては最大規模の惨事となった。（以下略）」

・東名高速道路日本坂トンネル車両火災の感想

「今回の事件で一番問題となるのは、なぜ6台の車の事故が173台もの車両の焼失に至ったのかという点だと思います。「火災進入禁止」の表示にも関わらず、多くの運転者が止まることなくトンネル内に入ってしまった。この時の心理として、高速道路で止まることは追突事故の危険が伴うため、多くの人がその恐怖に追い立てられていたのではないのか。（以下略）」

・エンロン不正会計問題の概要

「まずエンロン不正会計問題を調べるにあたって、エンロンという会社がどういうものかということから述べておく。エンロンは、1985年に石油や天然ガスのパイプラインを運営するアメリカの企業として発足。規制緩和の流れに乗って電力、天然ガスなどエネルギー関連製品をインターネットで取引する企業として急成長し、大企業となった。では、なぜエンロンは不正会計を行っていたのか？（以下略）」

・エンロン不正会計問題の感想

「このエンロン不正会計問題を調べていていろいろと感ずることがあった。こういうおおがかりな不正会計となると、エンロン社単独ではできないと考えそれについて調べてみるとさまざまな不正を隠すために政治献金を大量にばらまいていたという疑惑まであるそう。その相手はブッシュ大統領も含まれている。そうなってくると、この不正会計問題はアメリカの1つの企業の問題だけではなく、アメリカの組織全体に大きな疑問を感じる。（以下略）」

これらの概要と感想文に対して主な品詞の出現頻度を調べた。まず形態素解析にかける前に、被験者から提出されたレポートをテキストファイルの形で整理し、文章から不要なスペース、記号、タイトルなどは削除した。

雪印乳業食中毒事件の感想の最初の文を処理した例を以下に示す。左の数字が品詞の分類コード、中央が語の基本形、右が語の基本形の読みを表している。分類コードをもとにして、品詞の出現頻度などを処理することが可能となる。茶筌

形態素解析を用いた事故・リスクに関する記述の分析

では、同じ形容詞でも「形容詞－自立」、「形容詞－非自立」、「形容詞－接尾」のようにさらに細かく分類することができるが、本論文では、動詞、形容詞、副詞、連体詞、接続詞、助詞、助動詞、感動詞の大きな分類で処理をした。

57	この	コノ
9	雪印乳業	ユキジルシニユウギョウ
17	食中毒	ショクチュウドク
2	事件	ジケン
61	で	デ
14	私	ワタシ
61	が	ガ
51	恐ろしい	オソロシイ
47	思う	オモウ
21	こと	コト
65	は	ハ
79	,	,
2	老若男女	ロウニヤクナンニョ
14	誰	ダレ
65	も	モ
61	が	ガ
16	普段	フダン
2	口	クチ
61	に	ニ
47	する	スル
2	乳製品	ニユウセイヒン
63	によって	ニヨッテ
17	食中毒	ショクチュウドク
61	が	ガ
47	起きる	オキル
74	た	タ
62	と	ト
47	いう	イウ
21	こと	コト
74	だ	ダ
74	ある	アル
78	.	.

概要と感想文全てにおける品詞の出現頻度を表2に示す。品詞数をテキストの全文字数で割ったものを出現頻度として定義した。

表2

品詞	概要中の出現頻度	感想文中の出現頻度
動詞	6.70E-02	7.93E-02
形容詞	3.16E-03	6.21E-03
副詞	3.69E-03	8.98E-03
連体詞	4.55E-03	8.10E-03
接続詞	3.67E-03	5.14E-03
助詞	1.47E-01	1.69E-01
助動詞	4.11E-02	6.15E-02

70名の概要の総文字数は76968文字で、感想文の総文字数は48666文字であった。まずは、概要と感想文それぞれについて、品詞の出現頻度を比較した。図1から図7に各被験者の概要における頻度を横軸、感想文における頻度を縦軸にとって散布図として表示した。散布図としたので、概要と感想文における品詞の出現頻度の違いを集団として見る事が

図1 動詞の出現頻度

図2 形容詞の出現頻度

図3 副詞の出現頻度

図4 連体詞の出現頻度

図6 助詞の出現頻度

図5 接続詞の出現頻度

図7 助動詞の出現頻度

できる。図において点が集中する場合、個人によるばらつきが小さいことを意味する。また点が上方に集まる場合には、感想文において出現頻度が高いことを意味する。

動詞は、概要と感想文における出現頻度は同程度である。形容詞は、全体としては感想文の方が出現頻度が高くなるがばらつきが大きい。副詞は、感想文と概要中の出現頻度の差が最も大きかった。連体詞も、感想文の方が出現頻度が高くなっている。接続詞は、ばらつきが大きく、概要と感想文において顕著な差が認められない。助詞は、概要と感想文いずれにおいてもばらつきが小さい。助動詞は、全体としては感想文の方が出現頻度が高くなっている。感動詞の出現頻度は 4.6×10^{-5} と非常に小さく、数百字程度の文章の場合の定量的な議論は難しい。

続いて、出現頻度が高く、概要と感想文中の出現頻度の差が大きい副詞についてさらに解析を行った。表記が「特に」と「とくに」の場合のように、漢字の場合とかなの場合があるので、読みで統一して集計した。概要と感想文において、副詞の出現頻度のランキング上位を表3に示す。「さらに」は概要、感想文いずれにおいても出現頻度が高くなっている。

表3

概要	出現数	感想文	出現数
さらに	36	さらに	22
ほとんど	13	とくに	15
まず	11	もっと	14
まったく	10	しっかり	13
すぐ	9	どう	13
とくに	9	まったく	12
なぜ	9	なぜ	11
どうじに	8	そう	10
もともと	8	もし	10
きわめて	7	こう	9
じっさい	7	どうじに	9
じょじょに	7	まず	9
ほぼ	6	もう	9
ようやく	6	とうぜん	8
まだ	5	いかに	7
		じっさい	7
		とても	7
		ほんとうに	7
		より	7

形態素解析を用いた事故・リスクに関する記述の分析

概要では92種類、感想文では125種類の副詞が使われていた。感想文の方が副詞の種類が多いのは、副詞は、動詞や形容詞や他の副詞を修飾して、事態の様子・程度や話し手の気持ちを表す働きをする語類であるためだと考えられる[10, 11]。

3. tf-idfを用いた解析

続いて副詞に関してtf-idfを用いた解析を行った。tf-idfは、tf（単語の出現頻度）とidf（文章頻度の逆数の対数）の積であり、単語の重要度の指標として、情報検索やテキストマイニングの分野で用いられている[12-15]。

単語*i*のtf-idfは、以下のように定義される。

$$tfidf_i = tf_i \cdot idf_i \quad (式1)$$

$$tf_i = \frac{n_i}{\sum_k n_k}$$

$$idf_i = \log \frac{|D|}{|\{d : d \supset t_i\}|}$$

n_i : 単語*i*の出現頻度

$|D|$: 総文章数

$|\{d : d \supset t_i\}|$: 単語*i*を含む文章数

tf-idfでは、文章を特徴付ける単語をtfとidfという2つの特徴から重み付けをする。ひとつの文章中に同じ単語が多く出現すればtf-idfの値は大きくなり、多くの文章に単語が出現すればtf-idfの値は小さくなる。

表3に示した出現頻度の高い副詞（概要で15個、感想文で19個）についてtf-idfの値を求め、出現頻度（tf）との関係を調べた。図8に各副詞の概要と感想文における出現頻度を縦軸、tf-idfの値を横軸にとってプロットした。

図8の上部に2つ離れている点は概要、感想文共に出現個数のもっとも多かった副詞「さらに」である。この副詞を除くと、他の副詞は概要と感想文の2つのグループに分けられることが分かる。概要に比べ、感想文の方がtf、tf-idf共に

図8 副詞のtfとtf-idf

値が大きい傾向が明らかになった。特にtf-idfの値の差が顕著である。表3に示した副詞のうち「さらに」を除いたtf-idfの値の平均は、概要における値が 4.29×10^{-4} で、感想文における値が 7.62×10^{-4} であった。またこれらの副詞についてtf-idfの値のt検定を行った結果、t値は-10.51となり、P値は 2.1×10^{-11} となった。概要と感想文における「さらに」を除く副詞のtf-idfの値には、有意な差が見られた。つまり、概要と感想文を分類する手段として副詞のtf-idfを使用できる可能性があることが分かった。

4. ま と め

本研究では同じ著者の書いた文章でも、事故や企業トラブルなど、疑問、怒りなどを込めた感想文においては、概要を述べた文章に比べて品詞の出現頻度に明らかな違いが出ることが分かった。形態素解析で品詞の出現頻度を調べた結果、形容詞、副詞、連体詞の概要に対する感想文中の出現頻度の比はおよそ2倍の値となっていた。その中でも副詞の比率が最も大きかった。また感想文中において使用される副詞のバリエーションが増していることも分かった。概要と感想文における「さらに」を除く副詞のtf-idfの値には、有意な差が見られた。つまり、副詞のtf-idfを求めることは、概要と感想文を分類する手段として、使用できる可能性があることが分かった。この方法は、品詞の種類と頻度のみ注目しているので、分野にとらわれず、「事故」「心配」「危ない」などのキーワードがなくても、特定の感情を伴う文章を選別できる可能性を持っている。本手法は、以下のような応用が期待できる。

- 職員の定期的なレポートから職員の体調やストレスを予測し、事故を防止したり、離職率を下げる。
- カウンセリングの現場において、被験者の心理状態のモニタリングとして活用する。

一方で、品詞の種類と頻度の変化では、著者の心理状態の変化を捉えられても、具体的にどのような変化であるかを知ることは難しい。そこで本手法と、特定の名詞、形容詞、動詞などのキーワードに着目した分析を併用することが望ましい。

特定の副詞の出現頻度から特定の感情に関する文章を抽出できると結論付けるためには、今後より多くのデータを用いた統計的な議論が必要である。今回は概要と感想の文字数に差をつけて書いてもらったが、文字数が出現頻度の差に影響を及ぼす可能性もある。今後さらに条件を変えた評価が必要である。本論文では品詞の数を文字数で割って出現頻度の評価を行っているが、単語数に対する比率としての出現頻度の評価も必要である。また今回被験者が選んだ課題は、被験者が直接または身近に関わったケースはなく、新聞等の記者の意見の影響を受けている可能性がある。本人の個人的な感想文を扱うためにはより慎重な実験、分析が必要となる。また今回は主に品詞の量に関する評価を行ったが、特徴的な品詞の言語学的な意味にも注意した調査も必要である[16, 17]。

参 考 文 献

- [1] Jin, M., and Murakami, M.: *Author's Characteristic Writing Styles as Seen Through Their Use of Commas*, Behavior-metrika, 20(2), pp.63-76, 1993.
- [2] 村上征勝：真贋の科学，朝倉書店，1994.
- [3] 川原稔，河野浩之，長谷川利治：文献データベース情報検索に対するデータマイニング技術の適用，情報処理学会論文誌，39(4)，pp.878-887，1998.
- [4] 松尾豊，石塚満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム，人工知能学会論文誌，17(3)，pp.217-223，2002.
- [5] 林俊克：VACASによる感性商品開発（ファンデーションの開発事例），感性工学，2(1)，pp.25-27，2002.
- [6] 林俊克：Excelで学ぶテキストマイニング入門，オーム社，2002.
- [7] 石田基広：Rによるテキストマイニング入門，森北出版，2008.
- [8] 妙木裕，松本裕治，長尾眞：汎用日本語辞書および形態素解析システム，情報処理学会第42回全国大会予稿集，2001.
- [9] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸：形態素解析システム『茶釜』version 2.3.3 使用説明書，2003.
- [10] 渡辺実：副用語の研究，明治書院，1983.
- [11] 仁田義雄：副詞的表現の諸相，くろしお出版，2002.
- [12] 相澤彰子：語と文書の共起に基づく特徴度の数量的表現について，情報処理学会論文誌，41(12)，pp.3332-3343，2000.
- [13] 大森信行，岡村潤，森辰則，中川裕志：tf-idf法を用いた関連マニュアル群のハイパーテキスト化，情報処理学会研究報告，97(85)，pp.111-118，1997.
- [14] 馬場肇：Namazuシステムの構築と活用，ソフトバンククリエイティブ，2001.
- [15] 佐藤翔輔，林春男，牧紀男，井ノ口宗成：TFIDFを用いた災害・危機に関する言語資料体からのキーワード自動抽出手法の外的妥当性の検証，地域安全学会論文集，9，pp.65-74，2007.
- [16] 白川博之 監修：中上級を教える人のための日本語文法ハンドブック，スリーエーネットワーク，2001.
- [17] 寺村秀夫，鈴木泰，野田尚史，矢澤真人編：ケーススタディ日本文法，桜楓社，1987.

湊 淳（非会員）

1984年 京都大学理学部卒，1988年 東京大学大学院物理工学専攻修士修了，国立環境研究所勤務の後，現在茨城大学大学院理工学研究科応用粒子線科学専攻教授，博士（工学），超音波，光計測，感性工学などの研究に従事。

鈴木 慎吾（非会員）

2008年 茨城大学大学院理工学研究科応用粒子線科学専攻修了，現在ブラザー工業に勤務。

村上 雄太郎（非会員）

1980年 東京外国語大学外国語学部卒，現在茨城大学教授，博士（文学），感性和言語表現との関係を重視した対照言語学と外国語教授法研究に従事している。

伊多波 正徳（非会員）

2003年 茨城大学博士課程理工学研究科情報・システム修了，現在茨城大学講師，コンピュータを利用した物理教育に関する研究に従事。

小澤 哲（正会員）

1977年 東京教育大学大学院物理学専攻修了，理学博士，茨城大学大学院理工学研究科教授，専門は計算機シミュレーション。